A Data Mining Approach to Indirect Inference

Michael Creel*
Universitat Autònoma de Barcelona

Second version, October 2009

Abstract

Consider a model with parameter ϕ, and an auxiliary model with parameter θ. Let ϕ^h be a randomly sampled from a given density over the known parameter space. Monte Carlo methods can be used to draw simulated data and compute the corresponding estimate of θ, say $\tilde{\theta}_T^h$. A large set of tuples $\left(\phi^h, \tilde{\theta}_T^h\right)$ can be generated in this manner. Nonparametric methods may be use to fit the function $E\left(\phi|\tilde{\theta}_T = a\right)$, using these tuples. It is proposed to estimate ϕ using the fitted $E\left(\phi|\tilde{\theta}_T = \hat{\theta}_T\right)$, where $\hat{\theta}_T$ is the auxiliary estimate, using the real sample data. This is a consistent and asymptotically normally distributed estimator, under certain assumptions. Monte Carlo results for dynamic panel data and vector autoregressions show that this estimator can have very attractive small sample properties. Confidence intervals can be constructed using the quantiles of the ϕ^h for which $\tilde{\theta}_T^h$ is close to $\hat{\theta}_T$. Such confidence intervals are found to have very accurate coverage.

Keywords: simulation-based estimation; data mining; dynamic panel data; vector autoregression; bias reduction

JEL codes: C13, C14, C15, C33

1 Introduction

This paper presents a new simulation-based estimator. It is similar to the indirect inference (II) estimator (Gouriéroux, Monfort, Renault, 1993; Smith, 1993) in that it relies on an auxiliary estimator. The II estimator minimizes a measure of distance between the sample estimate of the parameter of an auxiliary model and the average of a number of replications of the auxiliary estimator, each computed using data generated by a trial value of the model's parameter. As such, the II estimator uses a double loop of minimizations: the inner loop where the auxiliary estimate is computed using data generated at a trial value of the model's parameter, and the outer loop where minimization is done over the model's parameter. The closely related efficient method of moments (EMM) estimator (Gallant and Tauchen, 1996) has an objective function that is based on the score of the

*Department of Economics and Economic History. michael.creel@uab.es. I thank S. Bonhomme for helpful comments. This work was supported by grants MICINN-ECO2009-11857 and SGR2009-578.

auxiliary model. The score function is always evaluated at the parameter estimate that results from using the sample data, but using simulated data generated at trial values of the model's parameter. As such, the auxiliary model is estimated only once, but the outer loop remains.

The proposal here is to sample from a chosen density over the parameter space, to use the draw on the parameter vector to generate a sample, and to use the sample to compute the value of an auxiliary estimator, in the manner of a Monte Carlo study. This process can be repeated to generate a very large set of pairs of parameter values and auxiliary estimates, a set as large as is desired. Then data mining methods can be used to learn about the relationship between auxiliary estimates and true parameter values. The method explored in this paper is to use nonparametric regression to estimate the expected value of the true parameter vector, conditional on the sample estimate of the auxiliary model's parameter vector. This expected value[1] is used as an estimator of the model's parameter vector.

Computing the expectation using nonparametric regression methods requires computing the estimate of the auxiliary model's parameter many times, using different trial values of the model's parameter, as is done in the inner loop of the II procedure. However, there is no outer loop. What are the possible advantages of this? As noted by Chernozukhov and Hong (2003), criteria functions of the form that defines the II estimator can have many local minima. If the the outer loop is avoided, then numeric difficulties will only be a problem when the auxiliary model is difficult to estimate. Another advantage is that the estimator can take advantage of complicated restrictions on the parameter space without having to impose such restrictions during minimization. The estimation of the parameters of a stationary vector autoregression in Section 5.2 gives an example. Another argument in favor of the proposed estimator is simply that seems to perform well in many cases, as is shown by example, below.

The next section introduces the estimator. In Section 3 it is shown to be consistent and asymptotically normally distributed. Section 4 discusses computation of confidence intervals. In Section 5, Monte Carlo work is presented for dynamic panel data models and vector autoregression models. Section 6 discusses extensions and conclusions.

2 The estimator

This section defines the proposed estimator. Because the estimator is similar to a certain type of indirect inference estimator that is proposed by Gouriéroux, Phillips and Yu (in press; henceforth referred to as GPY), I use a notation that closely follows that of their Section 3.

Suppose we have a model, indexed by a parameter $\phi \in \Phi \subset \mathbb{R}^k$. The model is simula-

[1] Because the precision of the nonparametric estimate of the expected value can be made as high as is desired simply be increasing the number of replications, I will refer to the estimated expected value simply as the expected value.

ble, so that given a parameter value ϕ, we can generate random samples of any size. The sample data, y_T, is a realized sample of size T, generated by the unknown true parameter value $\phi_0 \in \Phi$. Our problem is to estimate ϕ_0.

Consider an extremum estimator of the parameter of an auxiliary model:

$$\hat{\theta}_T = \arg\max_{\theta \in \Theta} Q_T(\theta; y_T), \tag{1}$$

where $Q_T(\theta; y_T)$ is an objective function.

Because the model is simulable, we are able to generate simulated samples that are of the same length as the observed sample. Suppose that the simulated true parameter value ϕ is chosen randomly by drawing from a chosen density f_ϕ that has support Φ. With this, one can draw a sample $\tilde{y}_T$ and then calculate the corresponding estimate, $\tilde{\theta}_T$. This may be repeated many times, to generate independently and identically distributed tuples $\left(\phi^h, \tilde{\theta}_T^h\right)$, $h = 1, 2, ..., H$. To illustrate, Figure 1 shows 200 such points. The true model used to generate these points is specified in Equation 5. The auxiliary estimator is the naive OLS estimator of θ in Equation 6, below. The Figure plots 200 pairs $\left(\phi^h, \tilde{\theta}_T^h\right)$, ignoring the other parameters (the α_i) of the true model.

When independent points are generated this way, one may consider the joint density $f_{\phi,\tilde{\theta}_T}$, the conditional density $f_{\phi|\tilde{\theta}_T=a}$, and the associated regression function $E\left(\phi|\tilde{\theta}_T = a\right)$, where a is some point in the range of the auxiliary estimator. The regression function $E\left(\phi|\tilde{\theta}_T = a\right)$ is similar to the binding function, defined by GPY as

$$b_T(\phi) = E(\tilde{\theta}_T^h(\phi)). \tag{2}$$

When ϕ and θ have the same dimension, the indirect inference estimator is $\hat{\phi}_T^{II} = b_T^{-1}(\hat{\theta}_T)$, and it can be computed by numerically solving $\hat{\theta}_T = E(\tilde{\theta}_T^h(\hat{\phi}_T^{II}))$. Usually, the analytic binding function is not known, but it can be calculated to any desired precision, by setting H as large as needed, using the simulated version

$$b_T^H(\phi) = \frac{1}{H}\sum_{h=1}^{H} \tilde{\theta}_T^h(\phi). \tag{3}$$

Contrarily, the regression function $E\left(\phi|\tilde{\theta}_T = a\right)$ cannot be learned by simple simulation, because we have no means of sampling ϕ while holding $\tilde{\theta}_T$ fixed at a given value. However, with some assumptions, it is possible to learn $E\left(\phi|\tilde{\theta}_T = a\right)$ up to any desired accuracy using nonparametric regression techniques. The H independent tuples $\left(\phi^h, \tilde{\theta}_T^h\right)$ discussed in the previous paragraph can be used to fit $E\left(\phi|\tilde{\theta}_T = a\right)$ nonparametrically. Likewise, the joint density $f_{\phi,\tilde{\theta}_T}$ can be fit using a kernel density estimator. To illustrate, the solid line in Figure 2 adds a kernel regression fit, based on $H = 500,000$ points, to the 200 points of Figure 1.

Assume that the chosen nonparametric estimator is uniformly consistent, as H increases. Kernel regression and density estimators are examples of nonparametric estima-

tors that have this property under reasonable assumptions. As the number of simulations, H, increases, the nonparametric fit will converge to $E\left(\phi|\tilde{\theta}_T = a\right)$. Furthermore, we may set H as large as is needed to obtain a fit to a given tolerance. Because computational power can render the difference between the nonparametric fit and the true expectation negligible, I simply abstract from the need to use a nonparametric estimator and henceforth, for the purposes of theory, treat the expectation as if it were a known function. The proposed estimator, say $\tilde{\phi}_T$, is the regression function, evaluated at the original estimate:

$$\tilde{\phi}_T = E\left(\phi|\tilde{\theta}_T = \hat{\theta}_T\right). \tag{4}$$

Henceforth, the estimator proposed here will be referred to as the "data-mining indirect (DMI) estimator".

The DMI estimator directly evaluates $E\left(\phi|\tilde{\theta}_T = \hat{\theta}_T\right)$, while the II estimator minimizes a measure of distance between $\hat{\theta}_T$ and $E(\tilde{\theta}_T^h(\hat{\phi}_T^{II}))$. There is a certain similarity in the basic concepts that define the two estimators. Figure 3 continues with the example described above that was used to create Figures 1 and 2, plotting $E\left(\phi|\tilde{\theta}_T = \hat{\theta}_T\right)$ and $b_T^{-1}(\hat{\theta}_T)$. Both of these functions are computed using kernel regression, using the 500,000 simulated points. The proposed estimator $\tilde{\phi}_T$ and the indirect inference estimator can be read off this Figure, given $\hat{\theta}_T$. Note that, conceptually, if we had a cloud made of an infinite number of points, $E\left(\phi|\tilde{\theta}_T = \hat{\theta}_T\right)$ corresponds to joining the expectations of points lying on vertical slices though the cloud (see Figure 1), while $b_T^{-1}(\hat{\theta}_T)$ joints the expectations of points on horizontal slices though the cloud. It is apparent in this Figure that $\tilde{\phi}_T$ and $\hat{\phi}_T^{II}$ are closely related, and are virtually identical for many values of the auxiliary parameter estimate. However, it is also clear that they are different estimators, at least for finite T, because the two lines diverge somewhat for certain values of the auxiliary estimator.

Note that the use of kernel smoothing to compute the inverse binding function, as is done here, imposes smoothness on the inverse binding function. The more usual procedure of numerically inverting the simulated binding function computed using Equation 3 does not impose smoothness, but instead relies on a large number of simulations to give smoothness as a result of uniform convergence in probability. It might be worthwhile to investigate the performance of the indirect inference estimator computed using the procedure suggested here.

3 Properties

In this section I show that the DMI estimator, $\tilde{\phi}_T$, defined in equation (4) is consistent and asymptotically normally distributed. I begin with assumptions:

Assumption 1. *$E\left(\phi|\tilde{\theta}_T = a\right)$ is continuous and bounded by an integrable function, $\forall T, \forall a \in \hat{\theta}_T(\Phi)$.*

Assumption 2. *$\lim_{T\to\infty} \hat{\theta}_T\left(\phi\right) = \theta_\infty(\phi)$, almost surely, $\forall \phi \in \Phi$*

Assumption 3. $\theta_\infty(\phi) : \Phi \to \theta_\infty(\Phi)$ *is injective.*

Assumption 1 is a simple regularity assumption. Likewise, Assumption 2 states that the auxiliary estimator is uniformly consistent for a pseudo-true value $\theta_\infty(\phi_0)$. Note that it may be biased and inconsistent for the true value, ϕ_0. The third assumption is perhaps the most fundamental of the above three assumptions. It is an identification assumption, so that knowledge of the almost sure limit of the auxiliary estimator gives us knowledge of the true parameter value. When both ϕ and θ are scalars, this assumption may be checked by visual inspection of a nonparametric fit to the binding function. The binding function must be strictly monotonic.

With these assumptions we can show that the proposed estimator is consistent:

Proposition 1. Given Assumptions 1, 2 and 3, $\tilde{\phi}_T \to^{a.s.} \phi_0$.
Proof: see Appendix.

To further clarify the relationship between the auxiliary estimator and DMI and the role of injectivity, consider a simple data generating process where the scalar parameter is known to lie in [0,1]. The auxiliary estimator is such that $\tilde{\theta}_T = 0.5 + \theta^2 + \frac{2\epsilon}{T}$ where $\epsilon \sim N(0,1)$. The auxiliary estimator is biased and inconsistent, but the relationship between the true parameter value and the pseudo-true value satisfies the injectivity assumption. Figures 5 and 6 plot several replications of the auxiliary and DMI estimators, for samples of size 10 and 50, respectively. We can see that the bias of the auxiliary estimator is largely eliminated. If this example is modified so that $\tilde{\theta}_T = 0.5 - 1.3\theta + \theta^2 + \frac{2\epsilon}{T}$, the injectivity assumption fails. Figure 7 shows the consequences.

Three additional assumptions are needed for asymptotic normality:

Assumption 4. $E\left(\phi|\tilde{\theta}_T = \theta_\infty(\phi_0)\right) = \phi_0 + o_p(T^{-1/2})$.

Assumption 5. $\sqrt{T}\left(\hat{\theta}_T - \theta_\infty(\phi_0)\right) \to^d N\left(0, V_\infty(\phi_0)\right)$

Assumption 6.

$$\left.\frac{\partial E\left(\phi|\tilde{\theta}_T = a\right)}{\partial a}\right|_{\theta^*} \to^{a.s.} G_\infty(\theta_\infty(\phi_0)),$$

a finite full rank matrix, for all θ^* that converge almost surely to $\theta_\infty(\phi_0)$.

Assumption 4 is a requirement that the function that defines the estimator, when evaluated at the pseudo-true value $\theta_\infty(\phi_0)$, must converge to the true parameter value sufficiently rapidly. I have not been able to prove that this results from simple fundamental assumptions, but simulations have been used to verify that it holds in a variety of cases. As an example, consider the case:

$$\begin{aligned} \theta_\infty(\phi) &= \phi + \phi^2 + \log(\phi + 1) \\ \hat{\theta}_T &= \theta_\infty(\phi) + \frac{\epsilon}{T} \\ \epsilon &\sim N(0,1) \end{aligned}$$

This case follows the above assumptions. $\hat{\theta}_T$ is biased and inconsistent for ϕ, but it is asymptotically normally distributed about $\phi + \phi^2 + \log(\phi + 1)$. For 20 different true values ϕ_0 evenly spaced on the interval [0,1], simulated $\tilde{\theta}_T^s$, $s = 1, 2, ..., 10^7$ were drawn, and $E\left(\phi | \tilde{\theta}_T = \theta_\infty(\phi_0)\right)$ was calculated by averaging the ϕ for which $\left|\tilde{\theta}_T^s = \theta_\infty(\phi_0)\right| < 10^{-5}$. Table 1 reports summary statistics for $\sqrt{T}\left(E\left(\phi | \tilde{\theta}_T = \theta_\infty(\phi_0)\right) - \phi_0\right)$, over the 20 true values of ϕ_0, for $T = 10^i$, $i = 1, 2., , , .5$. We can see that the mean is close to zero in all cases, and the standard deviation is declining to zero as T increases.

Assumption 5 simply states that the auxiliary estimator $\hat{\theta}_T$ is asymptotically normally distributed after centering about the pseudo-true value $\theta_\infty(\phi_0)$. This will hold for many types of auxiliary estimators and data generating processes.

Assumption 6 is an identification assumption: the auxiliary estimator must provide information about the parameter to be estimated.

Proposition 2. *Given Assumptions 4, 5 and 6,*

$$\sqrt{T}\left(\tilde{\phi}_T - \phi_0\right) \rightarrow^d N\left(0, G_\infty(\theta_\infty(\phi_0)) V_\infty(\theta_\infty(\phi_0)) G'_\infty(\theta_\infty(\phi_0))\right)$$

Proof: see Appendix.

4 Confidence intervals

The large number of tuples $\left(\phi^h, \tilde{\theta}_T^h\right)$, $h = 1, 2, ..., H$ that must be generated in order to compute the DMI estimator can be used to compute confidence intervals for the true parameter. For a given element of ϕ, say ϕ_i, the proposal is to choose a small resolution ϵ, and to select $A = \{\phi_i^h : \left|\tilde{\theta}_T^h - \hat{\theta}_T\right| < \epsilon\}$, where $\hat{\theta}_T$ is the realized sample value of the auxiliary estimator. The quantiles of A can be used to define limits of a confidence interval. For such a confidence interval to be accurate, the resolution ϵ must be small, and H must be large enough so that A contains many elements. An example that shows that confidence intervals computed in this way can be very accurate is given below in Section 5.1.1.

5 Monte Carlo results

5.1 Dynamic and nonlinear panel models

Dynamic and nonlinear panel models are important cases where econometric estimation methods often have a substantial bias. In this type of model, data have a double index: an observation is y_{it}, where $i = 1, 2, ..., N$ and $t = 0, 1, 2, ..., T$. Typically, there are N nuisance parameters. When T is small and fixed, which is a very relevant case empirically, the maximum likelihood estimator is inconsistent (Nickell, 1981). A number of estimators have been proposed to deal with this problem. Generalized method of moments/instrumental variables (GMM/IV) approaches include Holtz-Eakin, Newey and Rosen (1988),

Arellano and Bond (1991), Ahn and Schmidt (1995), Hahn (1997), Blundell and Bond (1998) and Alvarez and Arellano (2003). Another approach involves bias correction applied to the ML estimator. Examples include Bun and Caree (2005), Kiviet (1995) and Hahn and Kuersteiner (2002). Yet another approach is to use Bayesian priors as a means of reducing bias (Lancaster, 2002; Arellano and Bonhomme, 2008). Recently, Gouriéroux, Phillips and Yu (in press; GPY) propose an indirect inference estimator that uses the ML estimator to define binding functions.

In this sub-section, I consider several models that have been used in previous research, to facilitate comparison with other methods.

5.1.1 AR1 panel model

I use the Monte Carlo design of Hahn and Kuersteiner (2002), which was also used by GPY. Data are generated from the linear dynamic panel model

$$y_{it} = \alpha_i + \phi_0 y_{it-1} + \epsilon_{it} \tag{5}$$

where $\epsilon_{it} \sim N(0,1)$, $\alpha_i \sim N(0,1)$, $\phi_0 = 0$, 0.3, 0.6, 0.9 and α_i and ϵ_i are independently distributed. The initial condition is

$$y_{i0}|\alpha_i \sim N\left(\frac{\alpha_i}{1-\phi_0}, \frac{1}{1-\phi_0^2}\right).$$

Samples are generated for $N = 100$, 200 and $T = 5$, 10, 20.

I use $H = 500{,}000$ draws on $\left(\phi^h, \tilde{\theta}^h_{NT}\right)$ to be used as "data" for the nonparametric fitting process. These points were computed by drawing the autoregressive parameter from a uniform density on the stable region: $\phi^h \sim U(-1,1)$. For each ϕ^h, $\tilde{\theta}^h_{NT}$ is the ML ("fixed effect" or "within") estimator of ϕ^h. The 500,000 draws on $\left(\phi^h, \tilde{\theta}^h_{NT}\right)$ are specific to the sample size, N and T, but are not specific to the design point, $\phi_0 = 0$, 0.3, 0.6, 0.9.

Next, 5000 Monte Carlo samples are made for each design point $\phi_0 = 0$, 0.3, 0.6, 0.9, giving 5000 replications of the base estimator, $\hat{\theta}^j_T(\phi_0)$, $j = 1, 2, ..., 5000$. Finally, the 5000 replications of the base estimator are used to generate 5000 replications of the DMI estimator, using the kernel regression estimator

$$\tilde{\phi}^j(\phi_0) = \tilde{E}\left(\phi|\tilde{\theta}_T = \hat{\theta}^j_T(\phi_0)\right) = \frac{\sum_{h=1}^H \phi^h K\left(\frac{\tilde{\theta}^h_T - \hat{\theta}^j_T(\phi_0)}{\gamma}\right)}{\sum_{h=1}^H K\left(\frac{\tilde{\theta}^h_T - \hat{\theta}^j_T(\phi_0)}{\gamma}\right)},$$

$j = 1, 2, ..., 5000$, where $K(\cdot)$ is an Epanechnikov kernel, and the bandwidth γ is the rule-of-thumb[2] value $\gamma = H^{-1/5}$.

This same procedure is also applied using a naive OLS estimator based on the mis-

[2] See Li and Racine, 2007, pg. 66.

specified model

$$y_{it} = \theta y_{it-1} + u_{it} \tag{6}$$

as the starting point. This estimator simply ignores fixed effects and does not include a constant term. This naive estimator is very biased, but it has a small variance.

Table 2 presents the Monte Carlo results, along with the results for the indirect inference estimator proposed by GPY. In their Table 1, it is seen that the indirect inference estimator achieves a RMSE that is lower than that of the best of a set of competing estimators, for almost all designs. Comparing their Tables 1 and 2, it is seen that the indirect estimator with 250 simulated paths almost always achieves a lower bias than any of the competing estimators. For this reason, only the indirect estimator is used as a basis for comparison here. It is important to keep in mind that both the DMI and the II estimators require that the model be simulable, which in the present context means that these estimators require that the distribution of the α_i be known, up to parameters. Other estimators that do not have this requirement will be more generally applicable. When $T = 5$ or $T = 10$, the DMI estimator using the naive base estimator achieves the lowest RMSE, by a notable margin. The proposed estimator using the ML estimator as the base performs somewhat better than the indirect estimator when $T = 5$, and about the same when $T = 10$. For $T = 20$, the DMI estimator using the naive base estimator is dominated by the proposed estimator that uses the ML estimator as the base. The proposed estimator using the ML base and the indirect inference estimator have virtually identical RMSE's when $T = 20$.

Arellano and Bonhomme (2009) present some estimators for nonlinear panel data models that use robust priors to reduce bias. In some of their Monte Carlo work, they use the AR1 panel design of equation 7, with $N = 100$, $T = 10$, and the true parameter value $\phi_0 = 0.5$. They report results for a number of estimators. Of the feasible estimators they present, the one that achieves lowest mean squared error and lowest mean absolute error is their "robust, iterated ∞" estimator. Table 3 compares their results for their best estimator to the DMI estimator, computed as described above using $H = 500,000$ and 5000 Monte Carlo replications, using both the ML estimator and the naive OLS estimator as the base. The results for the DMI estimator are a little better than the Lancaster (2002) estimator, and a little worse than the Arellano-Bonhomme estimator. It should be emphasized that the DMI estimator requires that the model be fully simulable, which implies that the distribution of the individual effects is known. The Lancaster and Arellano-Bonhomme estimators do not have this requirement, so they are more generally applicable. However, these estimators require model-specific computations to be made in order to specify the prior information, while the method proposed here only requires simulations on the uncorrected ML or naive OLS estimators.

Confidence intervals The procedure for computing confidence intervals described in Section 4 was implemented using the $H = 500,000$ draws, and a resolution of $\epsilon = 0.0005$.

For each sample size and design point $\phi_0 = 0, 0.3, 0.6, 0.9$, and for each of the 5000 Monte Carlo replications, the above procedure is used to compute 90%, 95% and 99% confidence intervals, using the ML auxiliary estimator. The proportion of times that the true ϕ^h is contained in the confidence interval is computed. The results are in Table 4. The intervals are somewhat too broad for $T = 5$ and $\phi_0 = 0.9$, but in general, the confidence intervals are very reliable. Confidence intervals using the naive auxiliary estimator have very similar coverage, and for this reason are not reported.

5.1.2 AR1 panel model with incidental trend

GPY also present Monte Carlo results for an extension of the AR1 panel model, incorporating an incidental trend. The extended model is

$$y_{it} = \alpha_i + \beta_i t + \phi_0 y_{it-1} + \epsilon_{it} \tag{7}$$

where the design is the same as described following equation 5, except that $\alpha_i = \beta_i = 0$. The same procedure as described above was used to compute the DMI estimator, using only the ML estimator (see GPY, page 16 for formulae used to compute the ML estimator) as the base. In this case it would not be fair to use the naive estimator that ignores individual effects and trends as the base, because these are true restrictions on the model.

Table 5 gives the results. For $T = 5$, and $N = 100, 200$, neither the proposed estimator nor the indirect inference estimator dominates in terms of RMSE. The most notable difference is for $\phi_0 = 0.9$, where the proposed estimator has considerably lower RMSE, though it is more biased. For time series of length $T = 10$ or $T = 20$, DMI nearly always has RMSE lower than that of the indirect inference estimator, with the differences being most notable for $N = 100$.

5.1.3 Static logit panel model

The Arellano-Bonhomme (2009) estimator performed a little better than the proposed estimator in the case of the AR1 panel model, $\phi_0 = 0.5$ (see above). To further compare the approaches, their static logit Monte Carlo design (see their Section 9) is used here to evaluate the performance of the proposed estimator. The design of the experiment is

$$y_{it} = \mathbf{1}\left[x_{it}\phi_0 + \alpha_{i0} + \epsilon_{it} > 0\right]$$

where $x_{it} \sim N(0,1)$ and the individual effects $\alpha_{i0} \sim N(\bar{x}_i, 1)$, where $\bar{x}_i = \frac{1}{T}\sum_{t=1}^{T} x_{it}$. The ϵ_{it} are independent draws from the logistic CDF. The true value of $\phi_0 = 1$, and $N = 100$. The experiment is repeated for $T = 5, 10, 20, 100$. The base estimator is a quasi-maximum likelihood estimator that ignores the individual effects. That is, the base estimator is the estimator of the misspecified logit model that results from the above model, with the exception that, erroneously, it is assumed that $y_{it} = \mathbf{1}\left[x_{it}\phi_0 + \epsilon_{it} > 0\right]$. As

above, 500,000 draws of the base estimator are made, where ϕ was sampled from $U(0, 2)$. These 500,000 draws are used to compute five thousand Monte Carlo replications of the DMI estimator in the manner described above.

Table 6 presents results. These results may be compared with Tables 1 and 2 in Arellano and Bonhomme (2009). For the case of $T = 5$, the proposed estimator achieves bias, mean squared error (MSE) and mean absolute error (MAE) lower that those of any of the estimators considered by Arellano and Bonhomme. For larger sample sizes, the results start to become mixed, if we consider the criteria of bias, MSE and MAE. For $T = 10$, the DMI estimator outperforms all the estimators considered by Arellano and Bonhomme according to at least two of these three criteria. For $T = 20$, the DMI estimator has a performance similar to many of the estimators considered by Arellano and Bonhomme, except for the Lancaster (2002) estimator, which performs best. For $T = 100$, the proposed estimator is dominated by most of the alternatives.

The assumption that the distribution of the individual effects be known is quite implausible in this example. Nevertheless, the example serves to illustrate how the DMI estimator can achieve a good bias reduction in small samples, though use of a simple naive auxiliary model, when one is able to write a fully simulable model.

5.2 Vector autoregressions

It is well known that the OLS estimator of the parameter of an autoregressive model has a non-negligible small sample bias (Shaman and Stine, 1988). The problem extends to vector autoregressions (Nicholls and Pope, 1988; Pope, 1990), and it is a factor that contributes to the small sample bias of estimated impulse-response functions (Kilian, 1998). Here, I examine a simple stationary vector autoregressive model, and provide simulation results. In this experiment, the DMI estimator essentially removes small sample bias from coefficient estimates, and has a root mean squared error that is considerably smaller than that of the conditional maximum likelihood estimator.

The model under consideration is a 3 variable VAR: $y_t = (y_{t,1}, y_{t,2}, y_{t,3})$, where $y_t = Ay_{t-1} + \epsilon_t$ and $\epsilon_t \sim N(0, I_3)$. It is assumed that the system is stationary. The parameter space is defined as the set of A such that the elements on the main diagonal are between 0.3 and 1.3, and the elements off the main diagonal are between -0.5 and 0.5, plus the requirement that the eigenvalues of A lie within the complex unit circle, as is implied by stationarity.

The Monte Carlo simulations are done as follows. The elements of A are set initially set randomly following $a_{ij} \sim U(0.3, 1.3)$ if $i = j$ and $a_{ij} \sim U(-0.5, 0.5)$ if $i \neq j$. Then it is checked if the eigenvalues of A lie within the complex unit circle, and A a is rejected if this is not the case. If rejection occurs, a new trial is made until the stationarity requirement is satisfied. When an A that satisfies stationarity is found, a time series of length 130 is generated, and the first 100 observations are discarded. Finally, the coefficients of the model $y_{t,1} = \alpha + \beta_1 y_{t-1,1} + \beta_2 y_{t-1,2} + \beta_1 y_{t-1,3} + \epsilon_{t,1}$ are estimated by ordinary least

squares, which is the conditional maximum likelihood estimator, ignoring the restrictions on the parameter space. This was done to generate 100,000 replications. For each replication, we save the true coefficients that are in the first row of A, as well as the four estimated coefficient from the OLS regression.

Next, for each of the 100,000 replications, the DMI estimator was applied, for each of the three parameters A_{11}, A_{12}, and A_{13}. The kernel regression used to implement DMI conditioned on all all four parameter estimates of the OLS estimator. Table 7 contains the results. Note that the results in this table marginalize over the random design of A, which was described above. We can see that the OLS estimator has a substantial bias in the case of A_{11}, and that the DMI estimator essentially removes this bias. The DMI estimator has substantially smaller RMSE and MAE than does the OLS estimator, for all three parameters. It is worthy of note that the DMI estimator does not impose stationarity on the estimates, but that it is a weighted average of parameter values drawn from a parameter space that is restricted to contain only points that give a stationary model. The conventional OLS estimator ignores the stationarity restriction. Part of the better efficiency of the DMI estimator may be due to the fact that stationarity is taken into account, at least indirectly. This example also illustrates how complicated restrictions on the parameter space might be taken into account in other contexts.

To get an idea of how the results depend upon the true values of the parameters, Figure 8 plots the first 1000 errors in the own-autoregressive parameter, $\widehat{A_{11}} - A_{11}$ as a function of the true value, A_{11}, for both the OLS and DMI estimators. Note that the true values of A_{12} and A_{13} are not controlled for. We can see that the DMI estimator uniformly has a smaller variance, but there may be some conditional bias, especially for small values of A_{11}. Controlling for A_{12} and A_{13} could account for some of this bias, but it is probable that some of the bias is due to the fact that we are attempting to use the DMI estimator at the bounds of the parameter space. Kernel regression fitting near the limits of the data is known to suffer from bias, (Li and Racine, 2007, page 30) and no attempt has been made to control for this problem. To eliminate this effect, one could sample true parameter values from an artificially enlarged parameter space, so that conditioning points would be surrounded by neighbors in all directions.

6 Conclusions

The DMI estimator introduced in this paper requires a fully specified model, so that data can be generated by simulation. In common with other simulation-based estimators, its applicability is limited by this requirement. A second requirement is that there be a one-to-one relationship between the pseudo-true value of the auxiliary estimator and the true parameter. This requirement may be difficult to satisfy in some cases, if no consistent auxilary estimator is available. However, the examples given in the paper show that these requirements can be met in some cases, and that the DMI estimator can perform quite well in comparison to other applicable estimators. The paper also has presented a

means of computing confidence intervals that appears to work very well.

It is worth noting that the auxiliary model used to define the DMI estimator can be more complicated that the auxiliary estimators that have been used in the examples of this paper. If a given estimator is known to work well for a certain model, it could be used as the auxiliary estimator for DMI estimation.

If the dimension of the parameter vector is high, it may be computationally burdensome to use a large number of simulation draws and/or a fully nonparametric estimator such as kernel regression. One might choose to use informal approximation methods that can yield a reasonably good fit to the bias function while using a limited number of simulations. A reasonably high-order polynomial in the parameter of the auxiliary model could be used to define basis functions for a least squares fit. Once could contemplate using the estimator only for the parameters of most interest. Instead of using the entire estimated parameter vector of the auxiliary model as conditioning variables in the regression function, one could drop parameters that are suspected to have little effect on the parameter of interest, to reduce the dimension of the problem and thus save on computations. There are many such possibilities for economizing on computations. Given that kernel regression is a data parallel problem, one can overcome computational demands by using parallel computing techniques (Creel, 2005). The computational work reported in this paper was done using the GNU Octave programming language (`http://www.octave.org`), on a 32-core computational cluster made using the PelicanHPC distribution of GNU Linux (`http://pelicanhpc.org`). Use of PelicanHPC is very similar to what is described in Creel (2007). All software needed to replicate the results of this paper is available from the author.

Extensions to the method are not difficult to imagine. Figure 4 continues with the example discussed in Section 2, showing a kernel density plot of $f_{\phi,\tilde{\theta}_T}$, based on H =50,000 simulated points. Superimposed on the density is the line $E(\phi|\tilde{\theta})$, which defines the DMI estimator. It is apparent that the maximizer of the density, conditional on $\tilde{\theta}$, and the expectation $E(\phi|\tilde{\theta})$ are in general close to one another, for vertical slices though the Figure, but that they diverge somewhat when $\tilde{\theta}$ is close to one. One could use the maximizer of the density conditional on $\tilde{\theta}$ as an estimator of ϕ_0. One might also use the conditional median of ϕ given $\tilde{\theta}$. Perhaps an alternative such as these could have better efficiency than the DMI estimator proposed here.

7 Appendix: Proofs

Proof of Proposition 1

The estimator is

$$\tilde{\phi}_T \quad = \quad E\left(\phi|\tilde{\theta}_T\left(\phi\right) = \hat{\theta}_T\right).$$

By assumption 1, we may pass the limit operator though the expectation operator. With this, the condition that $\tilde{\theta}_T(\phi) = \hat{\theta}_T$ must hold in the limit. By Assumption 2,

$$\lim_{T\to\infty} \tilde{\theta}_T(\phi) = \theta_\infty(\phi)\, a.s.$$

and

$$\lim_{T\to\infty} \hat{\theta}_T = \theta_\infty(\phi_0),\ a.s.$$

Thus, in the limit, we must have

$$\theta_\infty(\phi) \quad = \quad \theta_\infty(\phi_0),$$

except on a set of probability zero. By Assumption 3, this can hold only if $\phi = \phi_0$. Thus, with probability one, $\lim \tilde{\phi}_T = E(\phi|\phi = \phi_0) = \phi_0$. ■

Proof of Proposition 2

The estimator is $\tilde{\phi}_T = E\left(\phi|\tilde{\theta}_T(\phi) = \hat{\theta}_T\right)$. A Taylor's series expansion about the pseudo-true value $\theta_\infty(\phi_0)$ gives

$$\tilde{\phi}_T \quad = \quad E\left(\phi|\tilde{\theta}_T(\phi) = \theta_\infty(\phi_0)\right) + \left.\frac{\partial E\left(\phi|\tilde{\theta}_T(\phi) = a\right)}{\partial a}\right|_{\theta^*} \left(\hat{\theta}_T - \theta_\infty(\phi_0)\right)$$

By Assumption 4, we can write

$$\sqrt{T}\left(\tilde{\phi}_T - \phi_0\right) = \left.\frac{\partial E\left(\phi|\tilde{\theta}_T(\phi) = a\right)}{\partial a}\right|_{\theta^*} \sqrt{T}\left(\hat{\theta}_T - \theta_\infty(\phi_0)\right)$$

Assumptions 5 and 6 give the result. ■

References

[1] Ahn, S. and P. Schmidt (1995) Efficient estimation of models for dynamic panel data, *Journal of Econometrics*, **68**, 5-27.

[2] Alvarez, J. and M. Arellano (2003) The time series and cross-section asymptotics of dynamic panel data estimators, *Econometrica*, **71**, 1121-1159.

[3] Arellano, M. and S. Bond (1991) Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations, *Review of Economic Studies*, **58**, 277-297.

[4] Arellano, M. and S. Bonhomme (2009) Robust priors in nonlinear panel data models, *Econometrica*, **77**, 489-536.

[5] Blundell, R. and S. Bond (1998) Initial Conditions and Moment Restrictions in Dynamic Panel Data Models, *Journal of Econometrics*, **87**, 115-143.

[6] Bun, M.J. and M.A. Carree (2005) Bias-corrected estimation in dynamic panel data models, *Journal of Business and Economic Statistics*, **23**, 200-210.

[7] Chernozhukov, V. and H. Hong (2003) An MCMC approach to classical estimation, *Journal of Econometrics*, **115**, 293-346.

[8] Creel, M. (2005) User-friendly parallel computations with econometric examples, *Computational Economics*, **26**, 107-128.

[9] Creel, M. (2007) I ran four million probits last night: HPC clustering with ParallelKnoppix, *Journal of Applied Econometrics*, **22**, 215-223.

[10] Gallant, A. R., and G. Tauchen (1996) Which moments to match?, *Econometric Theory*, **12**, 657-681.

[11] Gouriéroux, C., A. Monfort, and E. Renault (1993) Indirect inference, *Journal of Applied Econometrics*, **8**, S85—S118.

[12] Gouriéroux, Phillips and Yu (in press) Indirect inference for dynamic panel models, *Journal of Applied Econometrics.*

[13] Hahn, J. (1997) Efficient estimation of panel data models with sequential moment restrictions, *Journal of Econometrics*, **79**, 1-21.

[14] Hahn, J. and G. Kuersteiner (2002) Asymptotically unbiased inference for a dynamic model with fixed effects when both n and T are large, *Econometrica,* **70**, 1639-1657.

[15] Holtz-Eakin, D., W. Newey, and H. S. Rosen (1988) Estimating vector autoregressions with panel data, *Econometrica*, **56**, 1371-1395.

[16] Kilian, L. (1998) Small-sample confidence intervals for impulse response functions, *Review of Economics and Statistics,* **80**, 218–230.

[17] Kiviet, I., (1995) On bias, inconsistency and efficiency of various estimators in dynamic panel data models, *Journal of Econometrics*, **68**, 53-78.

[18] Lancaster, T. (2002) Orthogonal Parameters and Panel Data, *Review of Economic Studies*, **69**, 647-666.

[19] Li, Q. and J. Racine (2007), *Nonparametric econometrics*, Princeton University Press.

[20] Nicholls, D. F., and A. L. Pope (1988) Bias in the estimation of multivariate autoregressions, *Australian Journal of Statistics,* **30A**, 296–309.

[21] Nickell, S. (1981) Biases in dynamic models with fixed effects, *Econometrica*, **49**, 1417-1426.

[22] Pope, A.L., (1990) Biases of estimators in multivariate non-Gaussian autoregressions, *Journal of Time Series Analysis,* **11**, 249–258.

[23] Shaman, P. and R.A. Stine (1988) The bias of autoregressive coefficient estimators, *Journal of the American Statistical Association,* **83**, 842-848.

[24] Smith, A.A. (1993) Estimating nonlinear time-series models using simulated vector autoregressions, *Journal of Applied Econometrics*, **8**, S63-S84.

Figures

Figure 1: Simulated points $(\tilde{\theta}^h, \phi^h)$

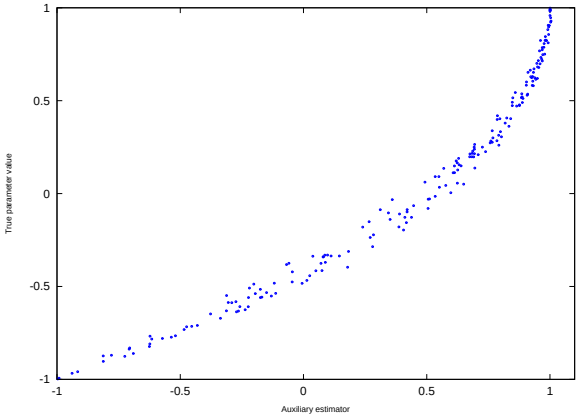


Figure 2: Kernel regression estimate of $E(\phi|\tilde{\theta})$

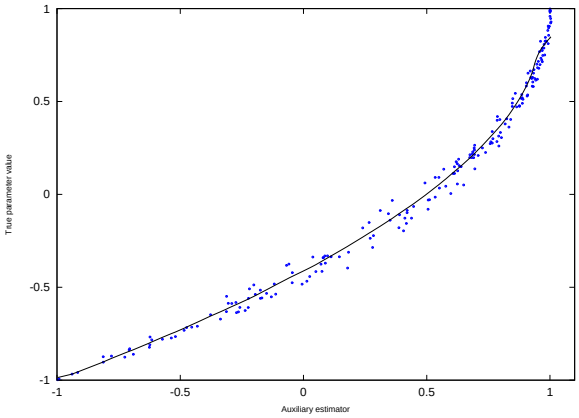


Figure 3: DMI=$E(\phi|\tilde{\theta})$ and II=$b^{-1}(\tilde{\theta})$

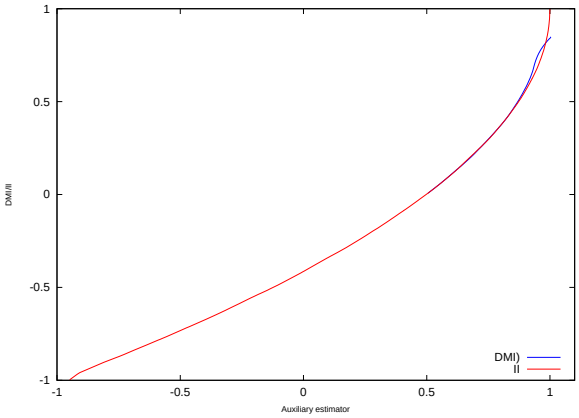


Figure 4: Kernel density estimate of $f_{\phi,\tilde{\theta}}$ with DMI=$E(\phi|\tilde{\theta})$ superimposed

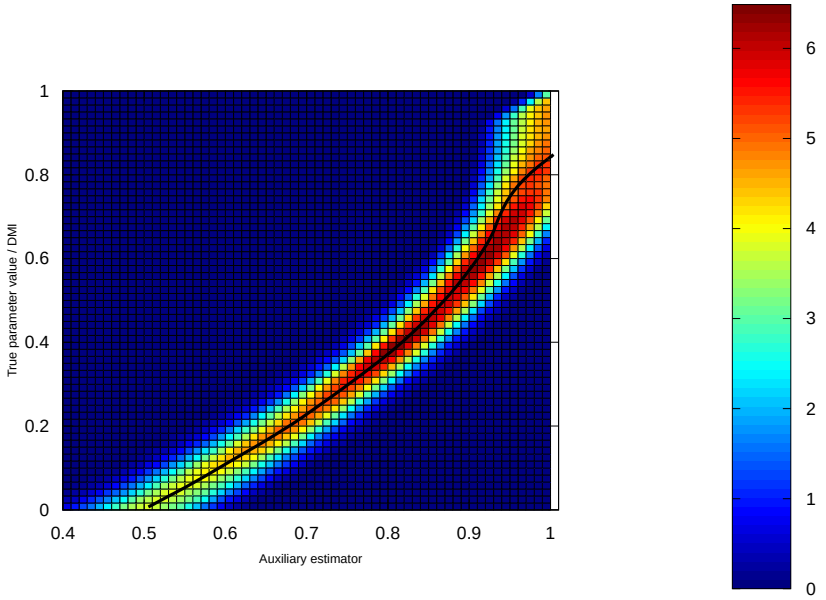


Figure 5: The relationship between the original estimator, AUX, and DMI. Samples of size $N = 10$.

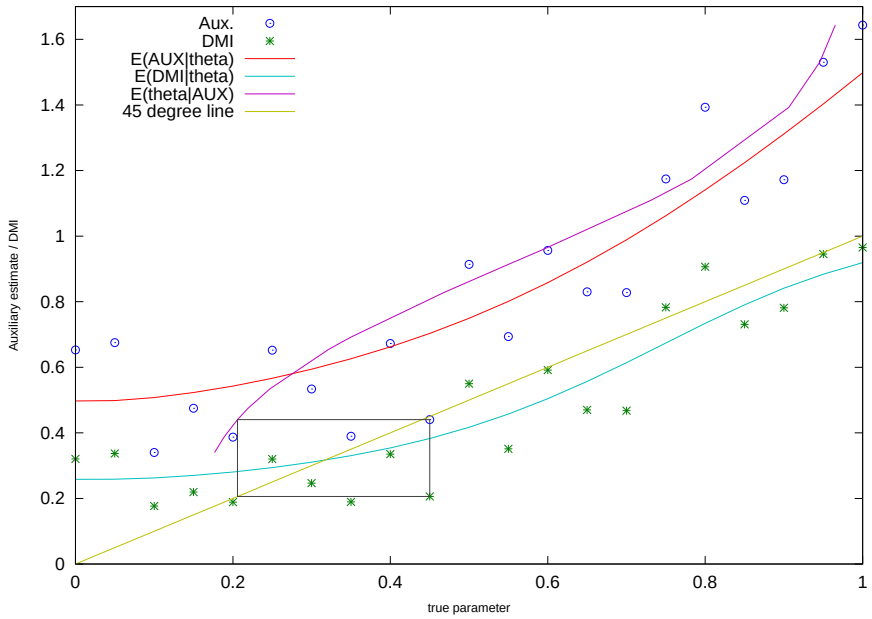


Figure 6: The relationship between the original estimator, AUX, and DMI. Samples of size $N = 50$.

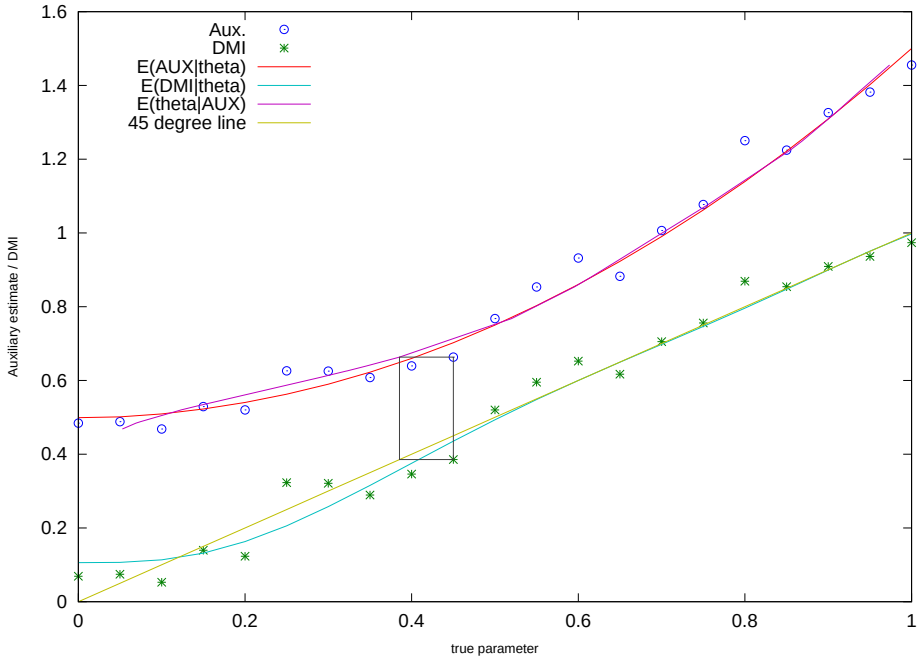


Figure 7: Consequences of failure of monotonicity.

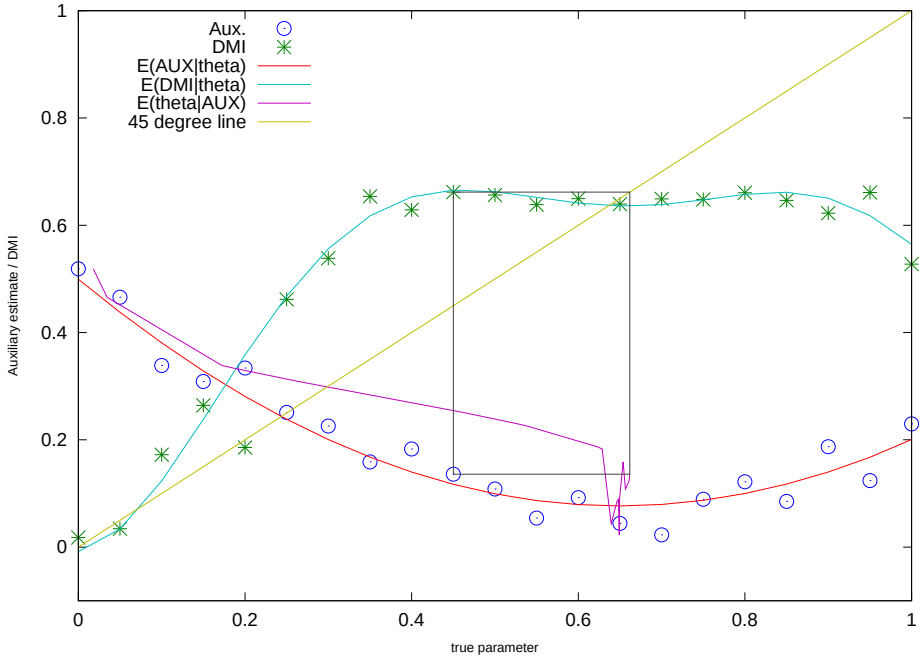
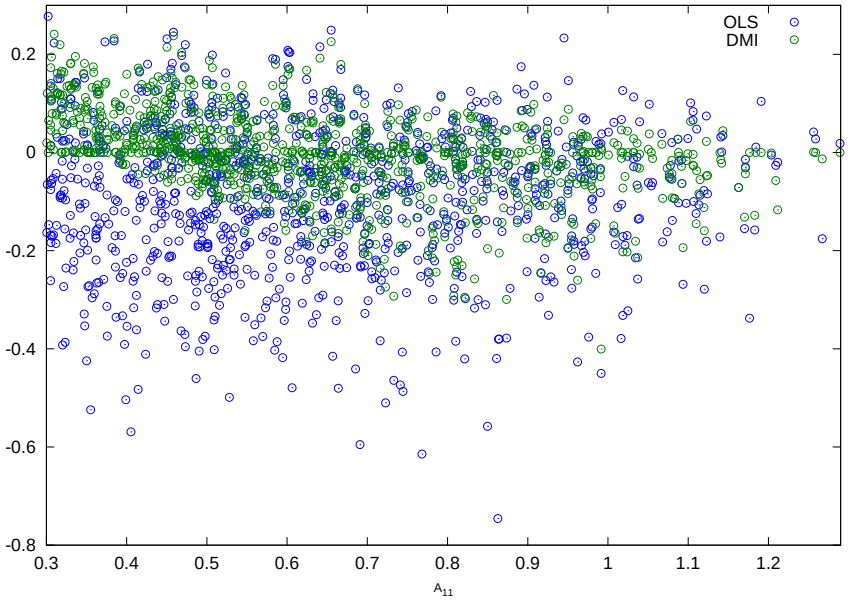


Figure 8: VAR model. $\widehat{A_{11}} - A_{11}$versus A_{11}. OLS and DMI.



Tables

Table 1: Checking assumption 4. Descriptive statistics , over 20 values of ϕ_0, of $\sqrt{T}\left(E\left(\phi|\tilde{\theta}_T = \theta_\infty(\phi_0)\right) - \phi_0\right)$

T	mean	st. dev.	min	max
10	0.002	0.024	-0.040	0.062
100	-0.001	0.005	-0.014	0.009
1000	-0.000	0.001	-0.002	0.002
10000	-0.000	0.000	-0.001	0.001
100000	0.000	0.000	-0.000	0.000

Table 2: Simple dynamic panel data model. Bias and RMSE of $\tilde{\phi}$ and indirect inference (II) estimator of Gouriéroux, Phillips and Yu, 2007. Source for II is Gouriéroux, Phillips and Yu, 2007, Table 2.

Case			Bias			RMSE		
T	N	ϕ	$\tilde{\phi}$ (ML)	$\tilde{\phi}$ (naive)	II	$\tilde{\phi}$ (ML)	$\tilde{\phi}$ (naive)	II
5	100	0.0	0.004	0.000	0.001	0.058	0.050	0.057
5	100	0.3	0.004	0.000	-0.001	0.064	0.045	0.081
5	100	0.6	0.004	0.011	0.000	0.069	0.045	0.070
5	100	0.9	-0.023	-0.035	0.000	0.057	0.036	0.076
5	200	0.0	0.001	0.000	0.000	0.041	0.035	0.041
5	200	0.3	0.002	0.000	-0.010	0.045	0.033	0.074
5	200	0.6	0.002	0.009	-0.000	0.049	0.031	0.050
5	200	0.9	-0.011	-0.034	-0.003	0.044	0.034	0.054
10	100	0.0	0.000	0.000	0.001	0.036	0.043	0.035
10	100	0.3	0.001	-0.000	0.000	0.036	0.039	0.036
10	100	0.6	-0.000	0.008	0.000	0.038	0.033	0.037
10	100	0.9	-0.004	-0.034	-0.001	0.034	0.034	0.040
10	200	0.0	0.000	-0.000	0.000	0.025	0.031	0.025
10	200	0.3	0.000	0.001	-0.000	0.026	0.028	0.026
10	200	0.6	0.001	0.009	0.000	0.027	0.024	0.026
10	200	0.9	-0.001	-0.033	0.002	0.026	0.033	0.028
20	100	0.0	0.000	0.001	0.001	0.024	0.040	0.024
20	100	0.3	-0.001	-0.000	0.001	0.023	0.035	0.024
20	100	0.6	-0.000	0.007	0.000	0.022	0.029	0.022
20	100	0.9	0.000	-0.033	0.000	0.020	0.033	0.021
20	200	0.0	-0.000	0.001	0.000	0.017	0.028	0.017
20	200	0.3	0.000	0.001	0.000	0.017	0.026	0.016
20	200	0.6	-0.000	0.008	0.000	0.015	0.021	0.015
20	200	0.9	0.001	-0.033	0.000	0.014	0.033	0.014

Table 3: Simple dynamic panel data model. $N = 100$,$T = 10$,$\phi_0 = 0.5$. "ML" is the maximum likelihood estimator without bias correction. "Lancaster" is the estimator proposed in Lancaster (2002), using Arellano and Bonhomme's equation 19. "AB" is the Arellano-Bonhomme robust, iterated ∞ estimator. Source for Lancaster and AB is Arellano and Bonhomme (2006) Table 3, page 35.

Estimator	Mean	Median	St. Dev.	MSE	MAE
ML	0.333	0.328	0.0320	0.0290	0.167
Lancaster	0.504	0.506	0.0374	0.00140	0.0302
AB	0.499	0.497	0.0323	0.00104	0.0264
$\tilde{\phi}$, ML	0.501	0.501	0.0368	0.00135	0.0292
$\tilde{\phi}$, naive	0.503	0.504	0.0331	0.00111	0.0267

Table 4: Confidence interval coverage. Simple dynamic panel data model. $\tilde{\phi}$ is computed using the ML auxiliary estimator.

Case			Coverage		
T	N	ϕ	90%	95%	99%
5	100	0.0	0.8974	0.9446	0.9900
5	100	0.3	0.9036	0.9492	0.9880
5	100	0.6	0.9036	0.9532	0.9886
5	100	0.9	0.9430	0.9736	0.9926
5	200	0.0	0.8960	0.9480	0.9872
5	200	0.3	0.8950	0.9450	0.9846
5	200	0.6	0.9060	0.9472	0.9886
5	200	0.9	0.9138	0.9646	0.9924
10	100	0.0	0.8970	0.9452	0.9870
10	100	0.3	0.9156	0.9552	0.9878
10	100	0.6	0.8928	0.9502	0.9890
10	100	0.9	0.9154	0.9592	0.9908
10	200	0.0	0.8856	0.9490	0.9898
10	200	0.3	0.8856	0.9518	0.9886
10	200	0.6	0.8940	0.9422	0.9884
10	200	0.9	0.8934	0.9470	0.9880
20	100	0.0	0.9016	0.9466	0.9884
20	100	0.3	0.8966	0.9514	0.9910
20	100	0.6	0.8952	0.9522	0.9890
20	100	0.9	0.8836	0.9382	0.9846
20	200	0.0	0.8992	0.9518	0.9864
20	200	0.3	0.8816	0.9444	0.9884
20	200	0.6	0.9028	0.9566	0.9900
20	200	0.9	0.9036	0.9604	0.9882

Table 5: Panel data model with incidental trend. Bias and RMSE of $\tilde{\phi}$ and indirect inference (II) estimator of Gouriéroux, Phillips and Yu, 2007. Source for II is Gouriéroux, Phillips and Yu, 2007, Table 3.

Case			Bias		RMSE	
T	N	ϕ	$\tilde{\phi}$	II	$\tilde{\phi}$	II
5	100	0.0	0.014	-0.019	0.089	0.078
5	100	0.3	0.038	-0.035	0.132	0.083
5	100	0.6	0.035	-0.037	0.133	0.151
5	100	0.9	-0.132	-0.050	0.158	0.252
5	200	0.0	0.006	-0.004	0.061	0.054
5	200	0.3	0.014	0.003	0.085	0.063
5	200	0.6	0.041	0.011	0.117	0.128
5	200	0.9	-0.094	-0.058	0.118	0.212
10	100	0.0	0.001	-0.034	0.040	0.054
10	100	0.3	0.002	-0.049	0.046	0.087
10	100	0.6	0.012	-0.034	0.064	0.068
10	100	0.9	-0.031	0.007	0.059	0.123
10	200	0.0	0.000	0.005	0.029	0.031
10	200	0.3	0.001	-0.010	0.033	0.081
10	200	0.6	0.005	0.010	0.043	0.041
10	200	0.9	-0.015	0.030	0.044	0.097
20	100	0.0	0.000	-0.008	0.025	0.027
20	100	0.3	0.001	-0.009	0.026	0.028
20	100	0.6	0.000	-0.010	0.028	0.030
20	100	0.9	0.004	-0.016	0.031	0.040
20	200	0.0	-0.000	0.000	0.018	0.018
20	200	0.3	0.000	0.000	0.018	0.019
20	200	0.6	0.000	0.001	0.019	0.020
20	200	0.9	0.008	0.010	0.026	0.032

Table 6: Static logit model, $\phi = 1$, $N = 100$,$T = 5$, 10, 20, 200.

T	Mean	Median	St. Dev.	MSE	MAE
5	1.013	1.007	0.137	0.019	0.110
10	1.008	1.005	0.095	0.009	0.076
20	1.005	1.005	0.067	0.004	0.054
100	1.002	1.002	0.035	0.001	0.028

Table 7: Vector autoregressive model

	$\widehat{A_{11}} - A_{11}$		$\widehat{A_{12}} - A_{12}$		$\widehat{A_{13}} - A_{13}$	
	OLS	DMI	OLS	DMI	OLS	DMI
Mean	-0.0932	-0.0004	0.0004	0.0000	-0.0004	-0.0000
Median	-0.0758	0.0000	0.0001	0.0000	-0.0002	0.0000
St. Dev.	0.1543	0.0884	0.1558	0.0960	0.1563	0.0957
RMSE	0.1802	0.0884	0.1558	0.0960	0.1563	0.0957
MAE	0.1363	0.0637	0.1191	0.0688	0.1196	0.0687